Port Elizabeth, South Africa, November 26-27, 2015

Using generalized maxout networks and phoneme mapping for low resource ASR– a case study on Flemish-Afrikaans

Reza Sahraeian[1], Dirk Van Compernolle[1] and Febe de Wet[2]

***Abstract*— Recently, multilingual deep neural networks (DNNs) have been successfully used to improve under-resourced speech recognizers. Common approaches use either a merged universal phoneme set based on the International Phonetic Alphabet (IPA) or a language specific phoneme set to train a multilingual DNN. In this paper, we investigate the effect of both knowledge-based and data-driven phoneme mapping on the multilingual DNN and its application to an under-resourced language. For the data-driven phoneme mapping we propose to use an approximation of Kullback Leibler Divergence (KLD) to generate a confusion matrix and find the best matching phonemes of the target language for each individual phoneme in the donor language. Moreover, we explore the use of recently proposed generalized maxout network in both multilingual and low resource monolingual scenarios. We evaluate the proposed phoneme mappings on a phoneme recognition task with both HMM/GMM and DNN systems with generalized maxout architecture where Flemish and Afrikaans are used as donor and under-resourced target languages respectively.**

***Index Terms*— Low resource ASR, phoneme mapping, Kullback Leibler Divergence, multilingual deep neural network.**

I. Introduction

Exploiting out-of-language data to develop high performance speech processing systems for low-resource languages has been extensively used recently [1][2]. However, sharing the knowledge across various languages is not a straightforward task because of differences such as different sets of subword units. In the literature, a common approach towards this is the creation of a universal phoneme set by first pooling the phoneme sets of different languages together and then merging them based on their similarity in both knowledge-based and data-driven fashions [3][4]. Knowledge-based phoneme mapping needs prior expert-knowledge of a phonetician and is an appropriate approach when we have no data for the target language. In practice, however, we usually have at least a few hours of data. To benefit from the available data, data-driven phoneme mapping can be used instead [5][6].

In the realm of multilingual neural networks [7], creating the target phoneme set for the multilingual training is commonly done (a) by joining of language-specific phoneme sets, (b) training neural networks where each language has its own output layer or (c) by mapping to a global phoneme set. The first two approaches have been successfully used when sufficient amount of training data for each language is available [8][9]. In the case of limited training data, however, using information from high resource language(s) by merging phoneme sets may be beneficial [10]. While the common approach for multilingual DNN training is that each language has its own output layer, our goal is to investigate if better performance can be gained by knowledge-based and data-driven phoneme mapping and which one performs best. This is a tricky issue as it depends on the languages. For example, if two languages are closely-related, IPA based mapping may work sufficiently well. Thus, in this paper, we conduct a case study for two related languages: Flemish and Afrikaans [12].

The data-driven approach we used is based on learning a phoneme mapping table by calculating KLD between pairs of phonemes in Flemish and Afrikaans. It is worth noting that similar works exist where a data-driven phoneme mapping is addressed by making the confusion matrix using multilingual neural networks [13][11]. However, the reported performance mostly degrades compared to the knowledge-based method. Moreover, there are two aspects in which this paper differs from [13]. First, the latter dealt with languages with moderate amounts of data and therefore DNN training where each language has its own output layer yields the best results; whereas, we deal with the resource-scarce target language and phoneme mapping is beneficial. Moreover, our approach is more flexible as we may assign more than one phoneme from Afrikaans to each phoneme of Flemish based on the confusion scores.

In addition, deep maxout networks have achieved improvements in various aspects of acoustic modelling for large vocabulary speech recognition systems including under-resourced and multilingual scenarios [14][15]. In this paper, we investigate the performance of state-of-the-art deep generalized maxout networks, [16], in the context of multilingual and under-resourced monolingual speech recognition.

This paper is organized as follows: in section II we describe deep generalized maxout network training. Then, the phoneme mapping issues for multilingual DNN and both the knowledge-based and data-driven approaches are explained in section III. The databases and the experiments are presented in section IV and V. Finally we present concluding remarks.

II. Deep Generalized Maxout Networks

A deep maxout neural network is simply a multilayer perceptron with many hidden layers before the softmax

*This work is based on research supported by the South African National Research Foundation as well as the fund for scientific research of Flanders (FWO) under project AMODA GA122.10N.

[1]Faculty of Electrical Engineering, KULeuven , 3001 Leuven, Belgium. Reza.Sahraeian@esat.kuleuven.be, Dirk.VanCompernolle@esat.kuleuven.be.

[2]HLT Research Group Meraka Institute, CSIR, South Africa. fdwet@csir.co.za

output layer and uses the maxout function to generate hidden activations [17]. Suppose $\mathbf{u}^{(l)} = [u_1^{(l)}, u_2^{(l)}, ..., u_I^{(l)}]$ is a set of activations in layer l; where

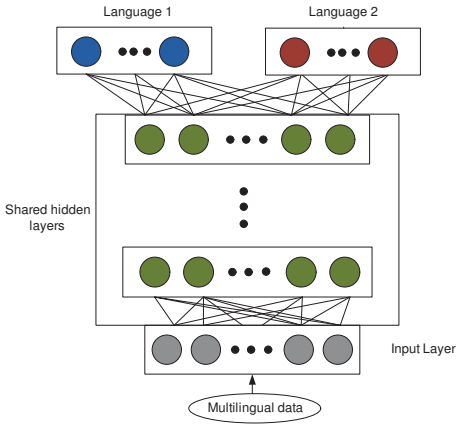
$$u_i^{(l)} = \max_j (h_j^{(l)}), \quad (i-1) \times g + 1 \leq j \leq i \times g \qquad (1)$$

The function takes the maximum over groups of inputs, $h_j^{(l)}$s, which are arranged in groups of g. $h_j^{(l)}$ is the jth element of $\mathbf{h}^{(l)} = \mathbf{W}^{(l)}\mathbf{u}^{(l-1)} + \mathbf{b}^{(l)}$. $\mathbf{W}^{(l)}$ is the matrix of connection weights between the $(l-1)$th and lth layers, $\mathbf{b}^{(l)}$ is the bias vector at the lth layer. In a maxout network, the nonlinearity is dimension-reducing and I is the dimensionality after the maxout function.

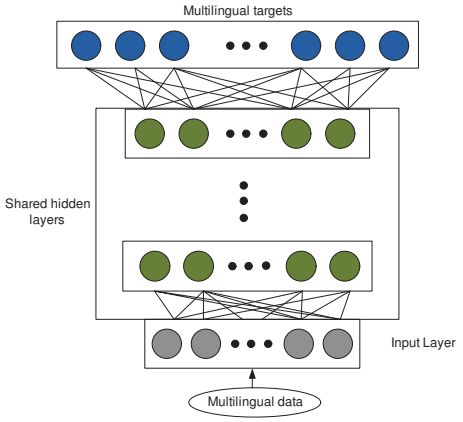
Generalized maxout networks may introduce other dimension reducing nonlinearities [16]. In this paper, we use the p-norm one:

$$u_i^{(l)} = (\sum_j |h_j^{(l)}|^p)^{\frac{1}{p}}, \quad (i-1) \times g + 1 \leq j \leq i \times g \qquad (2)$$

Where p is a configurable parameter. To train deep networks, greedy layer-wise supervised training [18] is used ; first, a randomly initialized network with one hidden layer is trained for a short time; then, the weights that go to the softmax layer are removed and a new hidden layer and two sets of randomly initialized weights are added. The neural network is trained again for the predefined number of iterations before a new hidden layer is inserted. This is repeated until we reach a desired number of layers. After the final iteration of training, the models from the last iterations are combined into a single model. In our study, the initial and final learning rates are specified by hand and equal to 0.02 and 0.004 respectively, and we always set $p = 2$. More details about the implementation and parameters are presented in [16].



(a) Multilingual DNN with language dependent output layer.



(b) Multilingual DNN with phoneme merged output layer.

Fig. 1. Multilingual DNNs with different types of output layers.

III. Phoneme Mapping in Multilingual DNN

Fig. 1 depicts the architecture of the typical multilingual DNN with shared hidden layers. In the multilingual target layer, each language can have its own output layer, Fig. 1-(a), or a common output layer is used Fig. 1-(b). In the latter, we need to provide a universal phoneme set; to this end, we may either consider a language label for each phoneme or merge phonemes. Simple concatenation of language specific phoneme sets, in the first scenario, may lead to performance degradation since very similar phones from different languages could be considered as different classes and the DNN would fail to discriminate between them [8]. For the second scenario, prior knowledge of a phonetician is required for the knowledge-based mapping which may not always be accurate and thus the DNN must encode disparate phonemes as a single class. This motivates us to investigate if a data-driven phoneme mapping can overcome the aforementioned problems. In the rest of this section, we describe the knowledge-based and data-driven phoneme mapping we used to train multilingual DNNs.

A. Knowledge-based Phoneme Mapping

The major assumption for knowledge-based (KB) phoneme mapping is that the articulatory representations of phonemes are similar and their acoustic realization can be assumed language independent. Based on this idea, universal phoneme inventories such as the IPA have been proposed [19]. In this study, the pronunciation dictionaries for the Afrikaans and Flemish include 37 and 47 phonemes respectively. In our KB phoneme mapping, each phoneme from the Flemish dictionary is mapped to only one of the phonemes in the Afrikaans one. To this end, 31 phonemes that share the same symbol in the IPA table are merged. However, there are 16 phonemes in Flemish without any IPA counterpart in Afrikaans which are mapped based on the linguistic knowledge. The phonemes: ɛ̃, ɑ̃, ɔ̃ and ʏ̃ are simply mapped to /ɛ n/, /ɑ n/, /ɔ n/ and /ʏ n/, and the rest are mapped as described in Table I.

B. Data-driven Phoneme Mapping

In our data-driven (DD) approach, we assume to have access to the pronunciation dictionary and the transcriptions for the target language. Then, each phoneme in Flemish can be mapped into N-best corresponding matches in the Afrikaans by calculating a confusion matrix.

Afterwards, a new pronunciation dictionary is created in which Flemish entries are described with the Afrikaans phonemes. Table II includes two examples explaining how

TABLE I

SUMMARY OF KNOWLEDGE-BASED PHONEME MAPPING BETWEEN FLEMISH(FL) AND AFRIKAANS(AFR) LANGUAGES.

Fl	Afr	Fl	Afr	Fl	Afr
ɣ	x	ʏː	ə	ɔː	ɔ
h	ɦ	oː	uə	ɛi	əi
ɪ	ɛ	eː	iə	ɛː	ɛ
ʏ	œ	aː	ɑː	ɑu	əu

the Flemish words "met" and "stipt" are phonetized in the original Flemish lexicon and the new KB and DD ones. In the first example, the phoneme "ɛ" in the Flemish is mostly confused with three phonemes in the Afrikaans: "ə", "œ" and "əi". Therefore, we consider three different pronunciations for this word based on the phoneme "ɛ" in the new lexicon. In this setup, the size of the new dictionaries increase rapidly with increasing N values. In addition, many of the Flemish phonemes have dominant matchings based on the confusion matrix; this is the case for almost all of the consonants. In this study, we set N=1 for the consonants and N=3 for the rest of the Flemish phonemes. It is also interesting to note that the Flemish phoneme "ɛ", for example, was merged with the Afrikaans phoneme of the same IPA symbol as in the KB phoneme mapping. However, "ɛ" is not among any of the three candidates chosen by DD approach. This indicates how differently the KB and the DD phoneme mapping may work.

In the second example, three different pronunciations for the word "stipt" are shown based on the phoneme "ɪ". This phoneme has no IPA matching in Afrikaans and is mapped to "ɛ" according to linguistic knowledge as shown in Table I. We should note that although the KB candidate for this phoneme is among those selected by DD approach, we have two more possible options for the mapping and depending on the context the best one will be chosen later based on the Viterbi alignment as a part of acoustic modeling. To

TABLE II

NEW PRONUNCIATION MODELING USING DD AND KB PHONEME MAPPING.

Fl word	Fl lexicon	DD lexicin	KB lexicon
met(1)	m ɛ t	m ə t	m ɛ t
met(2)	-	m œ t	-
met(3)	-	m əi t	-
stipt(1)	s t ɪ p t	s t ɛ p t	s t ɛ p t
stipt(2)	-	s t i p t	-
stipt(3)	-	s t ə p t	-

generate the confusion matrix, we measure the KLD between distributions of phonemes:

$$D(P \parallel Q) = \int P(x) log \frac{P(x)}{Q(x)} dx \tag{3}$$

Where P and Q represent density functions of the phonemes distributions in Afrikaans and Flemish respectively. It is worth noting that since KLD is not symmetric, it is normally appropriate for P to be the reference distribution and Q to be an approximation to it [20]. KLD is straightforward for normal distributions. However, for the multivariate Gaussian Mixtures Models (GMMs), the KLD is not analytically tractable and therefore we can use the variational approximation of KLD between GMMs [21]:

$$D^v(P \parallel Q) = \sum_a w_a log \frac{\sum_{a'} w_{a'} e^{-D(P_a \| P_{a'})}}{\sum_b \hat{w}_b e^{-D(P_a \| Q_b)}} \tag{4}$$

Where $P = \sum_a P_a$ and $P_a = w_a \mathcal{N}$, and $\mathcal{N}$ represents the normal distribution; similarly $Q = \sum_b Q_b$ and $Q_b = \hat{w}_b \mathcal{N}$. w and $\hat{w}$ are the Gaussian weights assigned to the Gaussian mixtures in the P and Q respectively. D^v is calculated for all pairs of phonemes in Afrikaans and Flemish to construct the confusion matrix. In this study, we use GMMs to model the phoneme distributions. Noting that the number of Gaussian components is set empirically and it equals two.

IV. DATABASES

A. Afrikaans data

The NCHLT corpus[1] [22] is an Afrikaans database including broadband speech sampled at 16 kHz. The phoneme set contains 37 phonemes and silence. We have been provided with a pronunciation dictionary as well as training, test and validation sets. All repeated utterances were removed from the original dataset. In our setting, to simulate a low resource condition, a data set including 1 hour of data and 188 speakers was extracted from the training part and used together with the original validation and test sets (Table III).

TABLE III

DESCRIPTION OF THE AFRIKAANS DATA SET AND A LOW RESOURCE SUBSET FOR TRAINING PURPOSES.

Set:	Train	Test	Dev
Duration	1h	2.2h	1.0h
# speakers	188	8	10

B. Flemish Data

The Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN) is a standard Dutch database that includes speech data collected from adults in the Netherlands and Flanders [23]. This dataset consists of 13 components that correspond to different socio-situational settings. In this study, we used Flemish data (audio recordings of speakers in Flanders) from component-o which contains read speech. This dataset includes 38 hours of speech sampled at 16KHz and we have taken 36h for the training and 2h for the evaluation. In this work, we used only the training part including 36 hours as donor data produced by 150 speakers.

[1]Available from the South African Resource Management Agency (http://rma.nwu.ac.za/).

Flemish words in the CGN pronunciation dictionary are phonetized by 47 phonemes which are mapped to the 37 phonemes of Afrikaans.

V. Experiments

This section describes the experimental study performed to evaluate the impact of deep generalized maxout networks for low resource ASR as well as the proposed phoneme mappings for multilingual DNN training. First, monolingual experiments on Afrikaans are presented which serves as a baseline. Then, we used Flemish to improve this performance in the context of multilingual DNN. In this study, we used the Kaldi ASR toolkit [24] for DNN training.

A. Monolingual Experiments

The first set of experiments was carried out on the Afrikaans language. We used a standard front-end by applying a Hamming window of 25ms length with a window overlap of 10ms. 13-dimensional features including 12 MFCC coefficients and the energy were extracted. Then, first and second derivatives were added and utterance-based mean and variance normalization was applied in both training and testing stages. These features were used to build 3-state left to right HMM triphone models with a total number of Gaussian components of $\sim$3000; this value was set using the validation set (Table III).

We trained a bi-gram phoneme model on the training set and the ASR performance is reported in phoneme error rate (PER). The neural network's inputs were the 24-dimensional FBANK features being concatenated with 7 left and 7 right neighbor frames, yielding a 360 dimensional input layer; then, an LDA transformation matrix was applied without dimensionality reduction. We observed that FBANK features outperform MFCCs as input features for DNN. In this set of experiments, we first trained standard DNN systems with *tanh* activation functions. The number of context-dependent triphone states (i.e. DNN targets) is 505; the number of units in each layer equals 100 to achieve the best results. Table IV provides the ASR performance using both HMM/GMM and the corresponding hybrid DNN systems. Since we have only one hour of training data, increasing the number of hidden layers may degrade the performance. The PERs for hybrid DNN systems with 1 and 2 layers are reported in Table IV; we observed higher PERs for more hidden layers. The best performance for monolingual DNN with *tanh* nonlinearity is obtained with one hidden layer.

TABLE IV
PER(%) for Afrikaans using HMM/GMM and hybrid DNN systems with *tanh* activation function trained on Afrikaans data only.

	HMM/GMM	Hybrid DNN	
		1 layer	2 layers
PER(%)	25.18	24.49	25.35

Then, we trained DNNs with the p-norm activation function; in this case, we have one more parameter which is the group size, g. The proper value for g and other neural network parameters such as number of hidden layers and the input dimensionality for the p-norm activation are jointly tuned on the validation set. In Table V the PERs for different numbers of hidden layers and different values of g are presented. In these experiments $I = 100$ and various input dimensionalities are investigated. Table V shows that the performance is improved when a generalized maxout network is used for such a low resource setting.

TABLE V
PER(%) on the Afrikaans using hybrid DNN systems with p-norm nonlinearity and various settings where the p-norm output dimensionality is $I = 400$.

input dim.	# of hidden layers			
	1	2	3	4
400	23.61	23.83	23.68	23.72
300	23.59	23.96	23.99	24.03
200	23.76	23.71	24.01	24.01

B. Multilingual Experiments

We subsequently merged the Flemish and Afrikaans training data based on both the knowledge-based and the data-driven universal phoneme sets explained in section III. Then, we trained a multilingual HMM/GMM system using 39-dimensional MFCC features. The numbers of tied-states used for the multilingual HMM/GMM system are 4131 and 3973 for the KB and DD approaches respectively.

Table VI gives the performance of the multilingual HMM/GMM systems for the two types of phoneme mapping by using the same bi-gram language model trained with 1 hour of Afrikaans. These results are presented here to evaluate the effectiveness of the DD phoneme mapping. As shown, DD phoneme mapping considerably improves the performance of multilingual HMM/GMM systems; yet, it can be seen that the PER is much higher than the monolingual case presented in Table IV and Table V.

TABLE VI
PER(%) comparisons for KB and DD phoneme mapping using a multilingual HMM/GMM system.

	KB mapping	DD mapping
PER(%)	45.89	39.81

Multilingual DNNs were subsequently trained by adopting context dependent decision trees and audio alignments from the multilingual HMM/GMM systems. In this set of experiments, the DNNs used p-norm activation functions and were trained from 15 consecutive frames and 24 FBANK features like DNN for monolingual setting. p-norm input and output dimensionality were empirically set to 1000 and 200 respectively. To bootstrap the acoustic model for Afrikaans, the hidden layers of the multilingual DNNs are shared and the softmax layer is replaced with the output layer corresponding to Afrikaans.

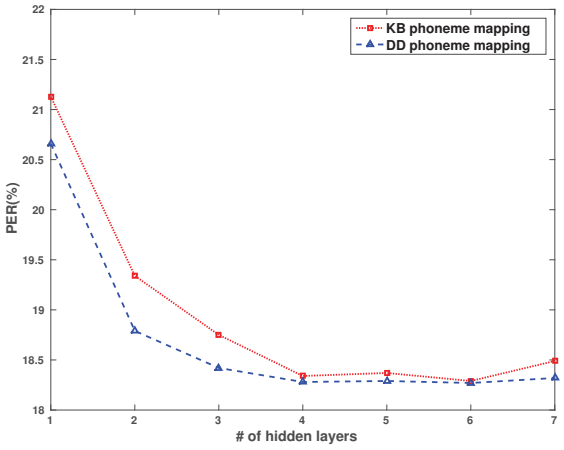


Fig. 2. PERs(%) comparisons for KB and DD phoneme mapping using multilingual DNN w.r.t. the number of hidden layers.

Fig. 2 shows a comparison of PERs obtained by multilingual DNNs with different numbers of hidden layers and reveals the following trends: first, both multilingual DNN systems provide significant reductions in ASR PERs when compared to the monolingual baseline systems presented in Table IV and Table V. Secondly, a comparison between the KB and DD phoneme mappings for DNN training shows that the ASR performance tends to improve in the case of using DD phoneme mapping. However, only marginal performance differences are observed if the neural networks are trained deep enough. This difference, however, depends on how similar the results of the two phoneme mapping techniques are. In this study, we observed that our DD technique maps all consonants to the same Afrikaans phonemes as the KB mapping does. Moreover, for many of the other Flemish phonemes, the selected KB candidate is among those chosen by the DD approach. For unrelated languages, however, DD phoneme mapping may perform differently and consequently lower PERs could be gained.

Finally, we examined another type of multilingual target where phoneme targets for Flemish and Afrikaans are kept separate Fig 1-(a). In this scenario, hidden layers are trained with data from both languages while the softmax layers are trained with language specific data where the number of output targets for Flemish is 4113 and 505 for Afrikaans.

TABLE VII
PER(%) FOR 6 HIDDEN LAYER MULTILINGUAL DNNS WITH AND WITHOUT PHONEME MAPPING.

	Phoneme mapping		No phoneme mapping
	KB	DD	
PER	18.29	18.25	21.04

Table VII shows that multileveled DNN approaches, either with or without phoneme mapping, improves ASR for low-resource languages. Moreover, we observe that phoneme mapping considerably improves the performance of multilingual DNNs. This can be due to the fact that Afrikaans and Flemish are closely related languages.

VI. CONCLUSION

This paper presented an investigation of using generalized maxout networks and phoneme mappings for multilingual DNN based acoustic modeling. Our aim was to improve a speech recognizer for Afrikaans (as an example of a resource-scarce language) with generalized maxout networks and by borrowing data from Flemish (as an example of a related well-resourced language). Phoneme sets of these two languages were merged in both knowledge-based and data-driven fashions. We proposed to use an approximation of KLD to generate the confusion matrix for the DD phoneme mapping. This DD approach led to a performance improvement which was more pronounced in the multilingual HMM/GMM system than the DNN one. Moreover, we observed that if we train neural networks deep enough, the performance difference between two phoneme mapping approaches decreases. We also observed that phoneme mapping is beneficial when Flemish data is used to boost the Afrikaans recognizer in the framework of the multilingual DNN.

REFERENCES

[1] D. Imseng, P. Motlicek, H. Bourlard and P. Garner, Using out-of-language data to improve an under-resourced speech recognizer, Speech Communication, vol. 56, 2014, pp. 142–151.
[2] L. Burget, et al., Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models, in Conf. Rec. 2010 IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP), pp. 4334–4337.
[3] V. B. Le, and L. Besacier, First Steps in Fast Acoustic Modeling for a New Target Language: Application to Vietnamese, in Conf. Rec. 2005 IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP), pp. 821–824.
[4] T. Schultz and A. Waibel, Language-independent and language-adaptive acoustic modeling for speech recognition, Speech Communication, vol. 35, 2001, pp. 31–51.
[5] K. C. Sim and H. Li, Robust phone set mapping using decision tree clustering for cross-lingual phone recognition, in Conf. Rec. 2008 IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP), pp. 4309–4312.
[6] W. Byrne, et al., Towards language independent acoustic modeling, in Conf. Rec. 2000 IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP), pp. II1029–II1032.
[7] J. T. Huang, J. Li,D. Yu, L. Deng and Y. Gong, Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers, in Conf. Rec. 2013 IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP), pp. 7304–7308.
[8] K. Veselỳ, M. Karafiát, F. Grézl, M. Janda and E. Egorova, The language-independent bottleneck features, in Conf. Rec. 2012 IEEE Workshop on Spoken Language Technology (SLT), pp. 336–341.
[9] S. Scanzio, P. Laface, L. Fissore, R. Gemello and F. Mana, On the use of a multilingual neural network front-end, in 2008 Proc. INTERSPEECH Conf., pp. 2711–2714.
[10] N. T. Vu, et al., Multilingual deep neural network based acoustic modeling for rapid language adaptation, in Conf. Rec. 2014 IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP), pp. 7639–7643.
[11] E. Egorova, K. Veselỳ, M. Karafiát, M. Janda and J. Cernocky, Manual and semi-automatic approaches to building a multilingual phoneme set, in Conf. Rec. 2013 IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP), pp. 7324–7328.
[12] W. Heeringa, and F. De Wet, The origin of Afrikaans pronunciation: a comparison to west Germanic languages and Dutch dialects, in 2008 Proc. Pattern Recognition Association of South Africa Conf., pp. 159–164.

[13] F. Grezl, M. Karafiát and M. Janda, Study of probabilistic and bottle-neck features in multilingual environment, in Conf. Rec. 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 359–364.
[14] P. Swietojanski, J. Li and J. T. Huang, Investigation of maxout networks for speech recognition, in Conf. Rec. 2014 IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP), pp. 7649–7653.
[15] Y. Miao, F. Metze, and S. Rawat, Deep maxout networks for low-resource speech recognition, in Conf. Rec. 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 398–403.
[16] X. Zhang, J. Trmal,D. Povey and S. Khudanpur, Improving deep neural network acoustic models using generalized maxout networks, in Conf. Rec. 2014 IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP), pp. 215–219.
[17] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville and Y. Bengio, Maxout networks, in 2013 Proc. ICML, pp. 1319–1327.
[18] Y. Bengio, P. Lamblin, D. Popovici and H. Larochelle, Greedy layer-wise training of deep networks, Advances in neural information processing systems, vol. 19, 2007, pp. 153–160.
[19] International Phonetic Association, Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet, Cambridge University Press, 1999.
[20] S. Kullback and R. A. Leibler, On information and sufficiency, The Annals of Mathematical Statistics, 1951, pp. 79–86.
[21] J. R. Hershey and P. A. Olsen, Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models, in Conf. Rec. 2007 IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP), pp. 317–320.
[22] E. Barnard, M. H. Davel, C. van Heerden, F. de Wet and J. Badenhorst, The NCHLT speech corpus of the South African languages, in 2014 Proc. Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU), pp. 194–200.
[23] N. Oostdijk, The Spoken Dutch Corpus. Overview and First Evaluation, in 2000 Proc. International Conference on Language Resources and Evaluation, pp. 887–894.
[24] D. Povey, et al., The Kaldi speech recognition toolkit, in Conf. Rec. 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 1–4.